Mario Baseball Data Analysis

Iszy Hirschtritt Licht

12/1/2020

Load Libraries

```r
library(tidyverse)
library(dplyr)
library(knitr)
library(weights)
library(scales)
library(stargazer)
library(ggthemes)
```

Load Data

```r
#Load Data
mario_data <- read.csv("Mario_Baseball_Data.csv")

#Clean Data
mario_data <- mario_data %>%
  replace(is.na(.), 0) %>%
  rename(
    date = Date,
    player_name = Player.Name,
    played_game = Games.Played,
    at_bats = AB,
    hits = Hits,
    runs_batted_in = RBI,
    homeruns = HR,
    stolen_bases = SB,
    special_hitting = Special,
    innings_pitched = IP,
    hits_allowed = Hits.1,
    runs_allowed = Runs,
    strikeouts = SO,
    big_plays = Big.Plays,
    special_pitching = Special.1,
    player_type = Player.Type,
    captain = Capitan) %>%
  mutate(date = as.Date(date,"%m.%d.%y"),
         played_game = as.factor(played_game),
         captain = as.factor(captain))
```

Hiting Data Analysis

```r
#Add Rate Data to Dataset
mario_data <- mario_data %>%
  group_by(player_name) %>%
  mutate(
    special_use_rate = sum(special_hitting)/sum(at_bats),
    batting_average = sum(hits)/sum(at_bats),
    era = (sum(runs_allowed)/sum(innings_pitched)*9),
    so9 = (sum(strikeouts)/sum(innings_pitched)*9),
    hip = sum(hits_allowed)/sum(innings_pitched))

#By Player Hitting
player_hitting <- mario_data %>%
  group_by(player_name) %>%
  summarise(batting_average = sum(hits)/sum(at_bats),
            special_use_rate = sum(special_hitting)/sum(at_bats))

kable(player_hitting, align = "lcc", col.names = c("Player", "Batting Average", "Special Use Rate"),
      digits = 3)
```

Player	Batting Average	Special Use Rate
Baby Bowser	0.147	0.059
Baby Luigi	0.227	0.045
Baby Mario	0.281	0.000
Birdo	0.345	0.258
Boo	0.356	0.000
Bowser	0.315	0.028
Daisy	0.346	0.151
Diddy Kong	0.154	0.154
DK	0.401	0.104
Drybones	0.286	0.005
Flying Goomba	0.000	0.000
Flying Koopa	0.324	0.006
Goomba	0.316	0.000
Grandpapa Toad	0.406	0.003
Hammer/Etc. Bro	0.361	0.000
King Boo	0.230	0.014
Koopa	0.303	0.011
Luigi	0.324	0.047
Magikoopa	0.224	0.006
Mario	0.455	0.452
Monty	0.194	0.000
Mumbo	0.221	0.006
Noki	0.234	0.065
Peach	0.248	0.096
Petey	0.295	0.000
Shy Guy	0.171	0.000
Toad	0.371	0.000
Toadette	0.214	0.000
Waluigi	0.321	0.161
Wario	0.200	0.080
Yoshi	0.333	0.130

```r
#By Player Type Hitting
player_type_hitting <- mario_data %>%
  group_by(player_type) %>%
  summarise(total_ab = sum(at_bats),
            total_hits = sum(hits),
            total_runs_batted_in = sum(runs_batted_in),
            total_homeruns = sum(homeruns),
            total_sb = sum(stolen_bases),
            batting_average = sum(hits)/sum(at_bats),
            special_use_rate = sum(special_hitting)/sum(at_bats),
            sb_hits = total_sb/total_hits)

kable(player_type_hitting, align = "lcccccccc", col.names = c("Player Type", "Total AB", "Total Hits",
                                                    "Total RBIS", "Total HR", "Total SB",
                                                    "Batting Average", "Special Use", "SB/Hits"),
      caption = "Hitting Stats by Player Type", digits = 3)
```

Table 2: Hitting Stats by Player Type

Player Type	Total AB	Total Hits	Total RBIS	Total HR	Total SB	Batting Average	Special Use	SB/Hits
Balance	2196	777	246	9	74	0.354	0.149	0.095
Power	1695	539	224	54	36	0.318	0.031	0.067
Speed	581	163	39	2	20	0.281	0.095	0.123
Technique	1901	614	189	5	48	0.323	0.031	0.078

```r
#By Y/N Captain Hitting
captain_stats <- mario_data %>%
  group_by(captain) %>%
  summarise(batting_average = sum(hits)/sum(at_bats))

kable(captain_stats, align = 'c', col.names = c("Captain Status", "Batting Average"),
      caption = "Are Captains Better Hitters?", digits = 3)
```

Table 3: Are Captains Better Hitters?

Captain Status	Batting Average
0	0.330
1	0.313

```r
#Running Batting Averages
mario_data <- mario_data %>%
  mutate(
    cum_at_bats = cumsum(at_bats),
    cum_hits = cumsum(hits),
    running_avg = cum_hits / cum_at_bats) %>%
  replace(is.na(.), 0)

#Plot Running Batting Averages
king_toad <- mario_data %>%
  filter(player_name == "Grandpapa Toad")
waluigi <- mario_data %>%
  filter(player_name == "Waluigi")
peach <- mario_data %>%
  filter(player_name == "Peach")
toad <- mario_data %>%
  filter(player_name == "Toad")
petey <-  mario_data %>%
  filter(player_name == "Petey")
mario <-  mario_data %>%
  filter(player_name == "Mario")

#Plot of 6 Players
ggplot() +
  geom_line(king_toad, mapping = aes(x = cum_at_bats,
                                     y = running_avg,
                                     color = "King Toad")) +
  geom_line(waluigi, mapping = aes(x = cum_at_bats,
                                   y = running_avg,
                                   color = "Waluigi")) +
  geom_line(peach, mapping = aes(x = cum_at_bats,
                                 y = running_avg,
                                 color = "Peach")) +
  geom_line(toad, mapping = aes(x = cum_at_bats,
                                y = running_avg,
                                color = "Toad")) +
  geom_line(petey, mapping = aes(x = cum_at_bats,
                                 y = running_avg,
                                 color = "Petey")) +
  geom_line(mario,mapping = aes(x = cum_at_bats,
                                y = running_avg,
                                color = "Mario")) +
  scale_x_continuous(breaks = pretty_breaks(n = 10)) +
  scale_y_continuous(breaks = pretty_breaks(n = 10)) +
  labs(title = "Running Batting Averages for Mario Baseball",
       subtitle = "Running Batting Average Across At Bats",
       x = "At Bats",
       y ="Batting Average") +
  scale_colour_discrete("Players") +
  theme_fivethirtyeight()
```
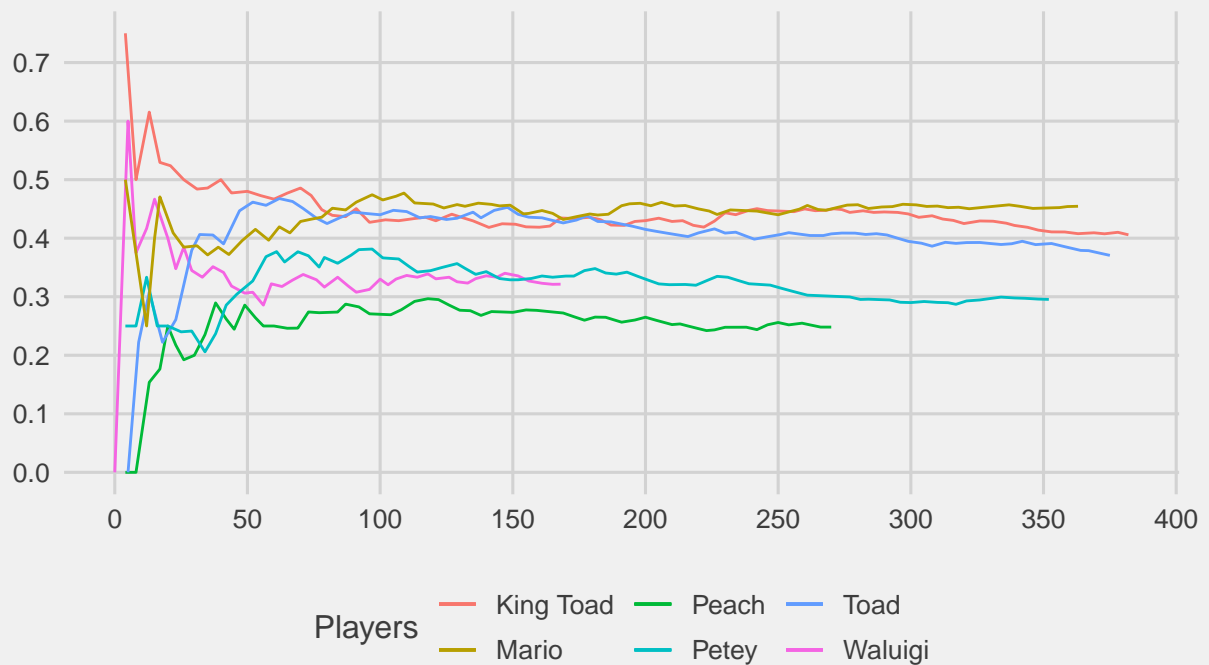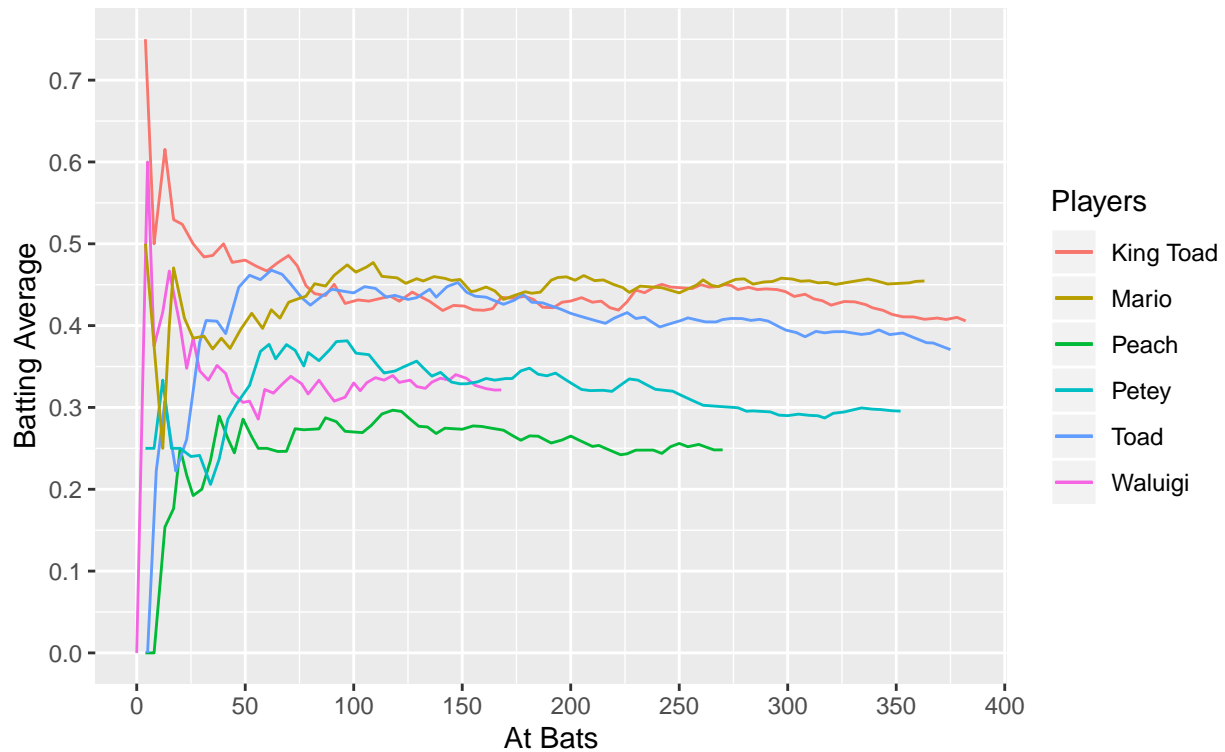
Running Batting Averages for Mario Baseball

Running Batting Average Across At Bats


```
ggplot() +
  geom_line(king_toad, mapping = aes(x = cum_at_bats, y = running_avg, color = "King Toad")) +
  geom_line(waluigi, mapping = aes(x = cum_at_bats, y = running_avg, color = "Waluigi")) +
  geom_line(peach, mapping = aes(x = cum_at_bats, y = running_avg, color = "Peach")) +
  geom_line(toad, mapping = aes(x = cum_at_bats, y = running_avg, color = "Toad")) +
  geom_line(petey, mapping = aes(x = cum_at_bats, y = running_avg, color = "Petey")) +
  geom_line(mario, mapping = aes(x = cum_at_bats, y = running_avg, color = "Mario")) +
  scale_x_continuous(breaks = pretty_breaks(n = 10)) +
  scale_y_continuous(breaks = pretty_breaks(n = 10)) +
  labs(title = "Running Batting Averages for Mario Baseball",
       subtitle = "Random Set of Player",
       x = "At Bats",
       y ="Batting Average") +
  scale_color_discrete("Players")
```

Running Batting Averages for Mario Baseball
Random Set of Player


```r
# Plot of all Players
ggplot(mario_data, aes(x=cum_at_bats,
                       y=running_avg,
                       group=player_name,
                       shape=player_name)) +
  geom_line() +
  facet_wrap(~ player_name) +
  labs(title = "Running Batting Averages for All Mario Baseball Players",
       x = "At Bats",
       y ="Batting Average")
```
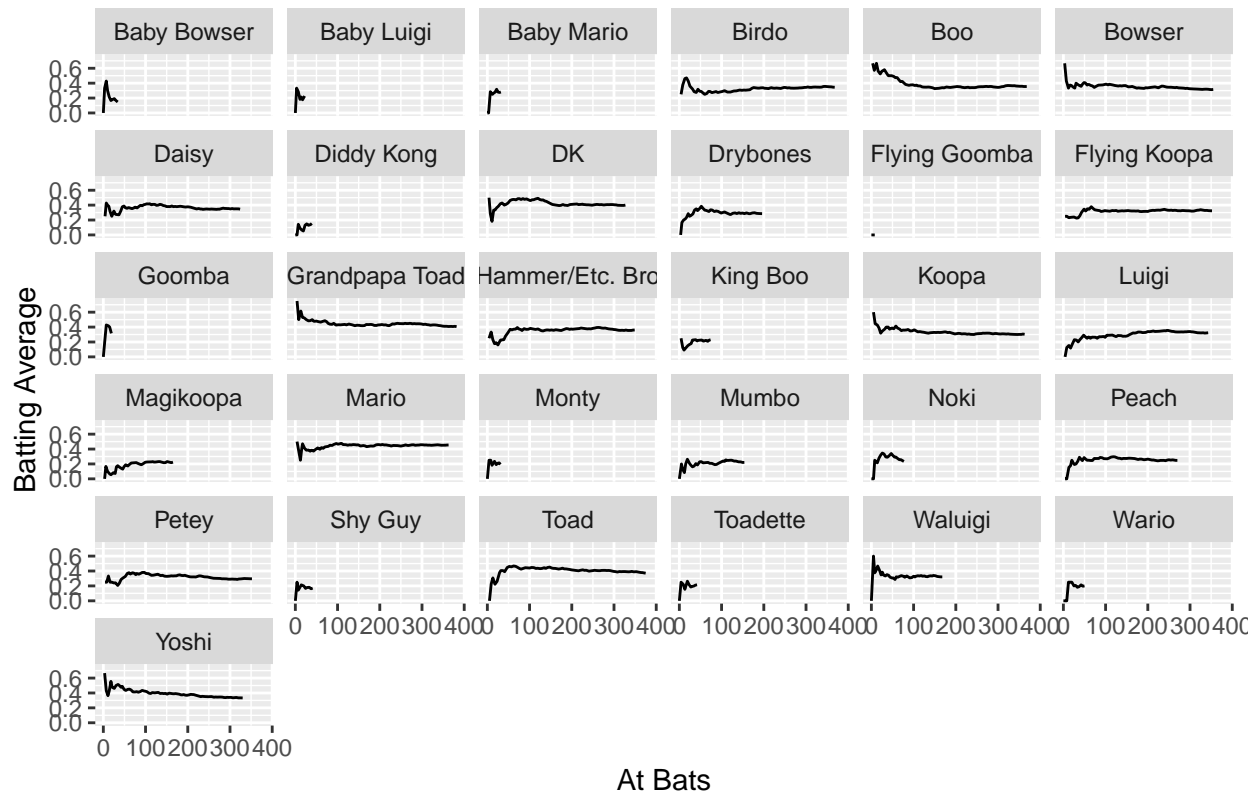
```
## Warning: The shape palette can deal with a maximum of 6 discrete values because
## more than 6 becomes difficult to discriminate; you have 31. Consider
## specifying shapes manually if you must have them.
```

Running Batting Averages for All Mario Baseball Players


At Bats

```
#Leaguewide running average
leaguewide_data <- mario_data %>%
  group_by(date) %>%
  summarise(
    total_hits = sum(hits),
    total_at_bats = sum(at_bats),
    total_average = sum(total_hits)/sum(total_at_bats)) %>%
  mutate(
    gameday = row_number(),
    cum_at_bats = cumsum(total_at_bats),
    cum_hits = cumsum(total_hits),
    running_avg = cum_hits / cum_at_bats)



#Leaguewide Average Plot
ggplot() +
  geom_line(leaguewide_data, mapping = aes(x=gameday, y=running_avg)) +
  labs(title = "Running Leaguewide Batting Average for Mario Baseball",
      x = "Gamedays",
      y ="Batting Average") +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10))
```


Running Leaguewide Batting Average for Mario Baseball
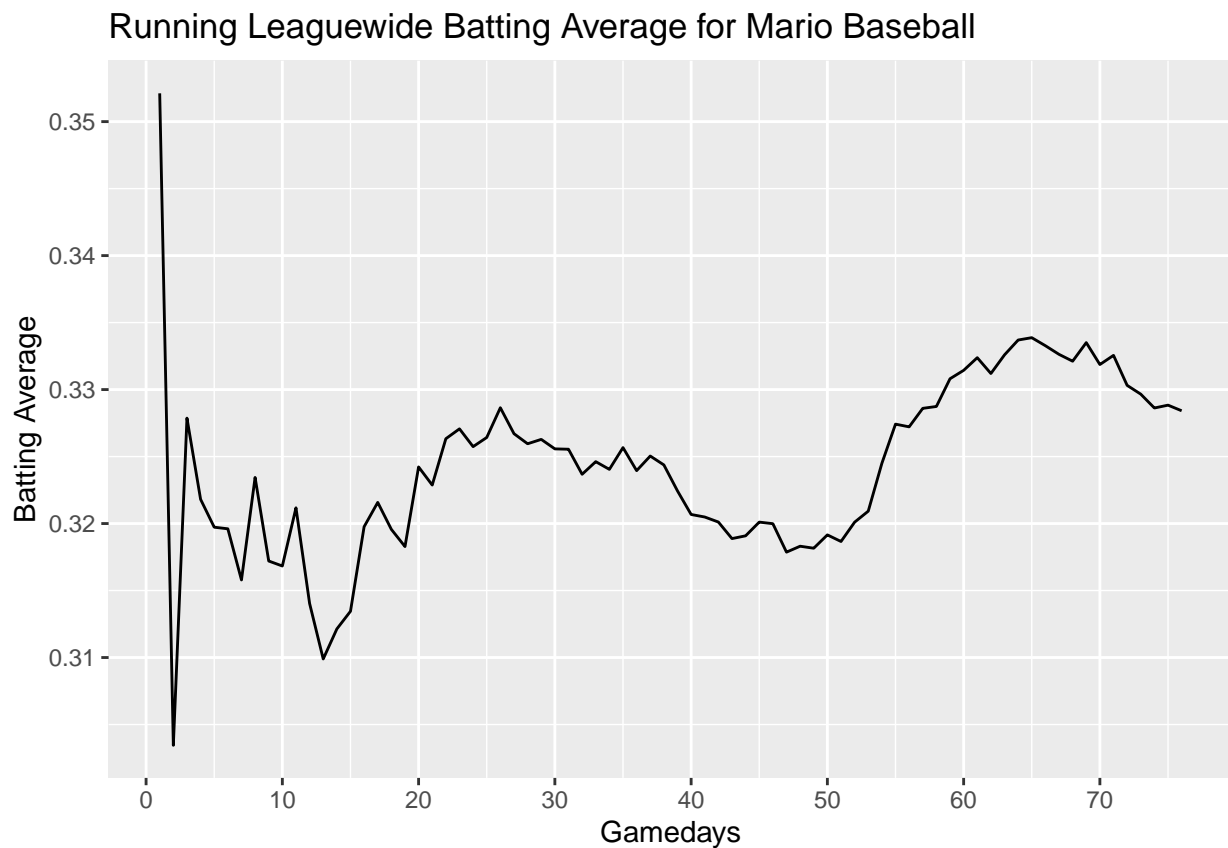
```
#Leaguewide Hits plot
ggplot(leaguewide_data, mapping = aes(x=gameday, y=total_hits)) +
  geom_point() +
  geom_smooth(method = "loess") +
    labs(title = "Running Leaguewide Hits for Mario Baseball",
      x = "Gamedays",
      y ="Total Hits")
```

Running Leaguewide Hits for Mario Baseball
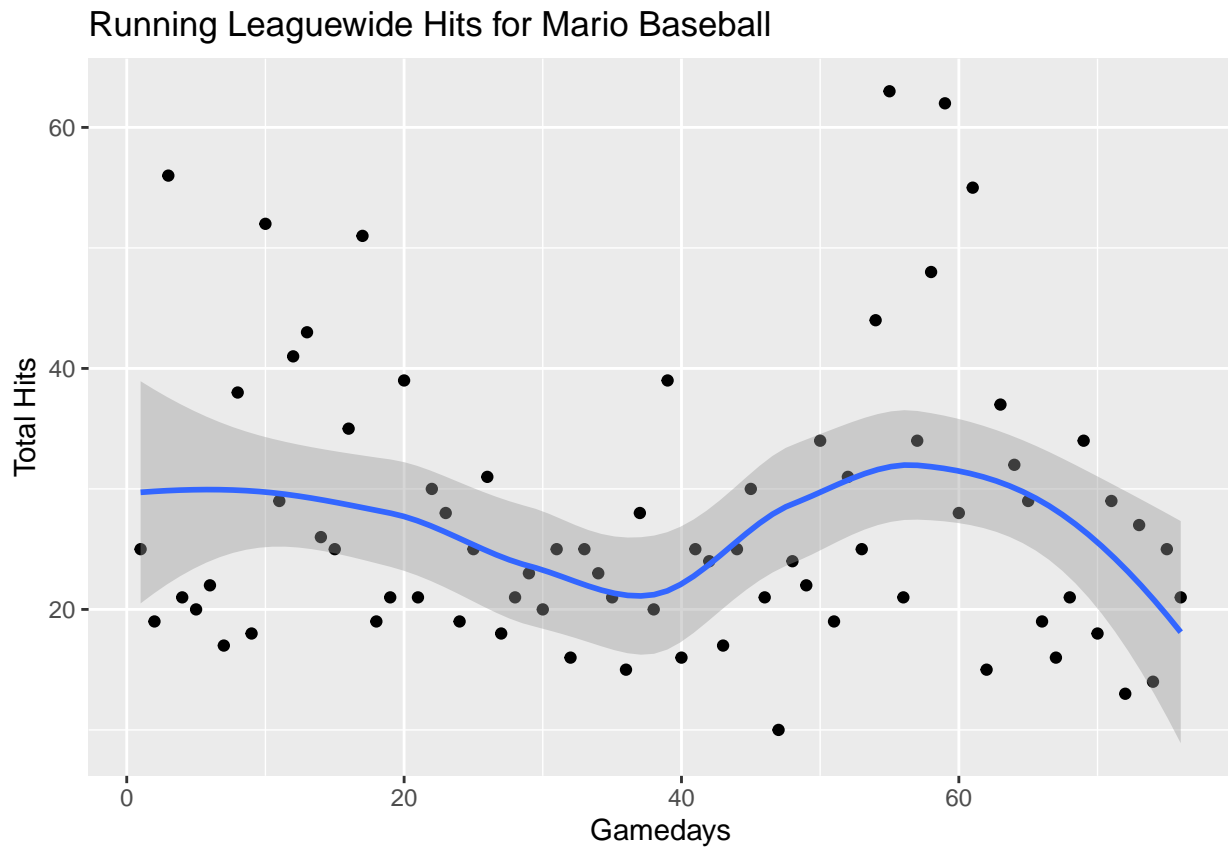
```
#Sample Regression
batting_avg_captain <- lm(batting_average ~ player_type + special_use_rate + captain ,
                     data = mario_data)

stargazer(batting_avg_captain,
          type = "latex", header = FALSE,
          title = "Regression of Batting Average on Player Type with Controls",
          intercept.bottom = FALSE, single.row=TRUE)
```

Table 4: Regression of Batting Average on Player Type with Controls

	Dependent variable:
	batting_average
Constant	0.296^{***} (0.003)
player_typePower	-0.035^{***} (0.004)
player_typeSpeed	-0.105^{***} (0.004)
player_typeTechnique	0.002 (0.004)
special_use_rate	0.291^{***} (0.016)
captain1	-0.010^{*} (0.005)
Observations	2,666
R^2	0.353
Adjusted R^2	0.351
Residual Std. Error	0.074 (df = 2660)
F Statistic	289.861^{***} (df = 5; 2660)
Note:	*p<0.1; **p<0.05; ***p<0.01

Pitching Data Analysis

```
#Player Pitching
player_era_1 <- mario_data %>%
  filter(sum(innings_pitched) >= 100) %>%
  group_by(player_name) %>%
  summarise(era = (sum(runs_allowed)/sum(innings_pitched)*9))

kable(player_era_1, align = "lc", col.names = c("Player", "ERA"),
      caption = "ERA for pitchers with 100+ Innings Pitched",
      digits = 3)
```

Table 5: ERA for pitchers with 100+ Innings Pitched

Player	ERA
Boo	4.335
Flying Koopa	3.970
Koopa	3.364
Waluigi	5.139

```
player_era_2 <- mario_data %>%
  filter(sum(innings_pitched) >= 25 & sum(innings_pitched) < 100) %>%
  group_by(player_name) %>%
  summarise(era = (sum(runs_allowed)/sum(innings_pitched)*9))

kable(player_era_2, align = "lc", col.names = c("Player", "ERA"),
      caption = "ERA for pitchers with 25-100 Innings Pitched",
      digits = 3)
```

Table 6: ERA for pitchers with 25-100 Innings Pitched

Player	ERA
Baby Luigi	2.979
Diddy Kong	2.306
DK	6.000

```r
#Player Type Pitching
player_type_pitching <- mario_data %>%
  group_by(player_type) %>%
  summarise(
    total_innings = sum(innings_pitched),
    era = (sum(runs_allowed)/sum(innings_pitched)*9),
    total_strikeouts = sum(strikeouts),
    total_big_plays = sum(big_plays))

kable(player_type_pitching, align = "lcccc",
      col.names = c("Player Type", "Innings", "ERA", "Strikeouts", "Big Plays"),
      caption = "Pitching and Fielding Stats by Player Type",
      digits = 3)
```

Table 7: Pitching and Fielding Stats by Player Type

Player Type	Innings	ERA	Strikeouts	Big Plays
Balance	574.587	3.634	522	172
Power	45.980	7.634	33	97
Speed	102.930	2.711	66	45
Technique	830.550	4.465	701	74

```r
#Running Pitching Stats
mario_data <- mario_data %>%
  mutate(
    cum_runs_allowed = cumsum(runs_allowed),
    cum_innings = cumsum(innings_pitched),
    running_era = (cum_runs_allowed / cum_innings)*9) %>%
  replace(is.na(.), 0)

waluigi <- mario_data %>%
  filter(player_name == "Waluigi")

flying_koopa <- mario_data %>%
  filter(player_name == "Flying Koopa")

koopa <- mario_data %>%
  filter(player_name == "Koopa")

boo <- mario_data %>%
  filter(player_name == "Boo")


#Plot Running Pitching
ggplot() +
  geom_line(koopa, mapping = aes(x = cum_innings, y = running_era, color = "Koopa")) +
  geom_line(boo, mapping = aes(x = cum_innings, y = running_era, color = "Boo")) +
  geom_line(flying_koopa, mapping = aes(x = cum_innings, y = running_era, color = "Flying Koopa")) +
  geom_line(waluigi, mapping = aes(x = cum_innings, y = running_era, color = "Waluigi")) +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  labs(title = "Running ERA for Mario Baseball",
       subtitle = "Players with 100+ Innings Pitched",
       x = "Innings Pitched",
       y ="ERA") +
  scale_colour_discrete("Players")
```
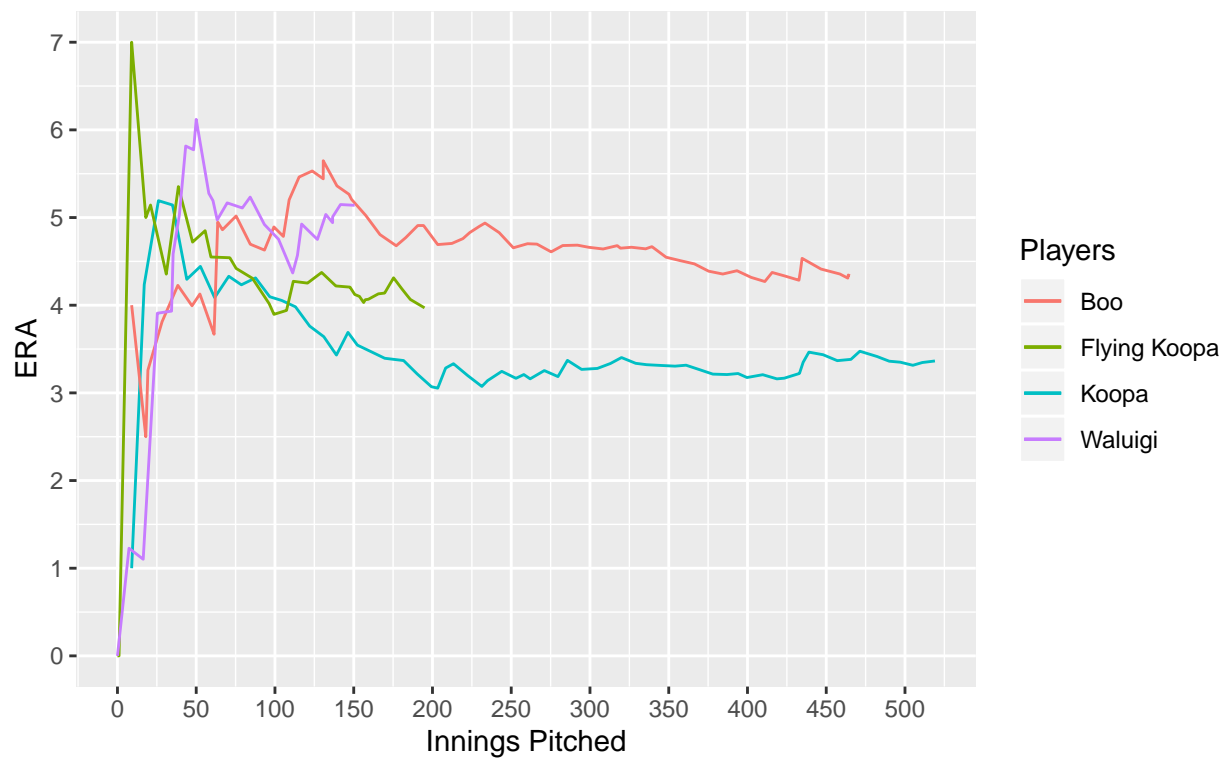
Running ERA for Mario Baseball

Players with 100+ Innings Pitched

```r
#Leaguewide Pitching
leaguewide_pitching <- mario_data %>%
  group_by(date) %>%
  summarise(
    total_innings = sum(innings_pitched),
    total_hits_allowed = sum(hits_allowed),
    total_runs_allowed = sum(runs_allowed),
    total_strikeouts = sum(strikeouts),
    total_era = ((sum(runs_allowed)/sum(innings_pitched))*9)
    ) %>%
  mutate(
    gameday = row_number(),
    cum_innings = cumsum(total_innings),
    cum_runs_allowed = cumsum(total_runs_allowed),
    cum_strikeouts = cumsum(total_strikeouts),
    running_so_9 = ((cum_strikeouts/cum_innings)*9),
    running_era = ((cum_runs_allowed / cum_innings)*9))


#Leaguewide ERA Plot
ggplot() +
  geom_line(leaguewide_pitching, mapping = aes(x=gameday,
                                               y=running_era,
                                               color = "ERA")) +
  geom_line(leaguewide_pitching, mapping = aes(x=gameday,
                                               y=running_so_9,
                                               color = "SO/9")) +
  scale_y_continuous("running_era", sec.axis = sec_axis(~ . * 1, name = "running_so_9")) +
  labs(title = "Running Pitching Stats for Mario Baseball",
       subtitle = "ERA and SO/9",
       x = "Gameday") +
    scale_colour_discrete("Pitching Stat")
```

Running Pitching Stats for Mario Baseball
ERA and SO/9
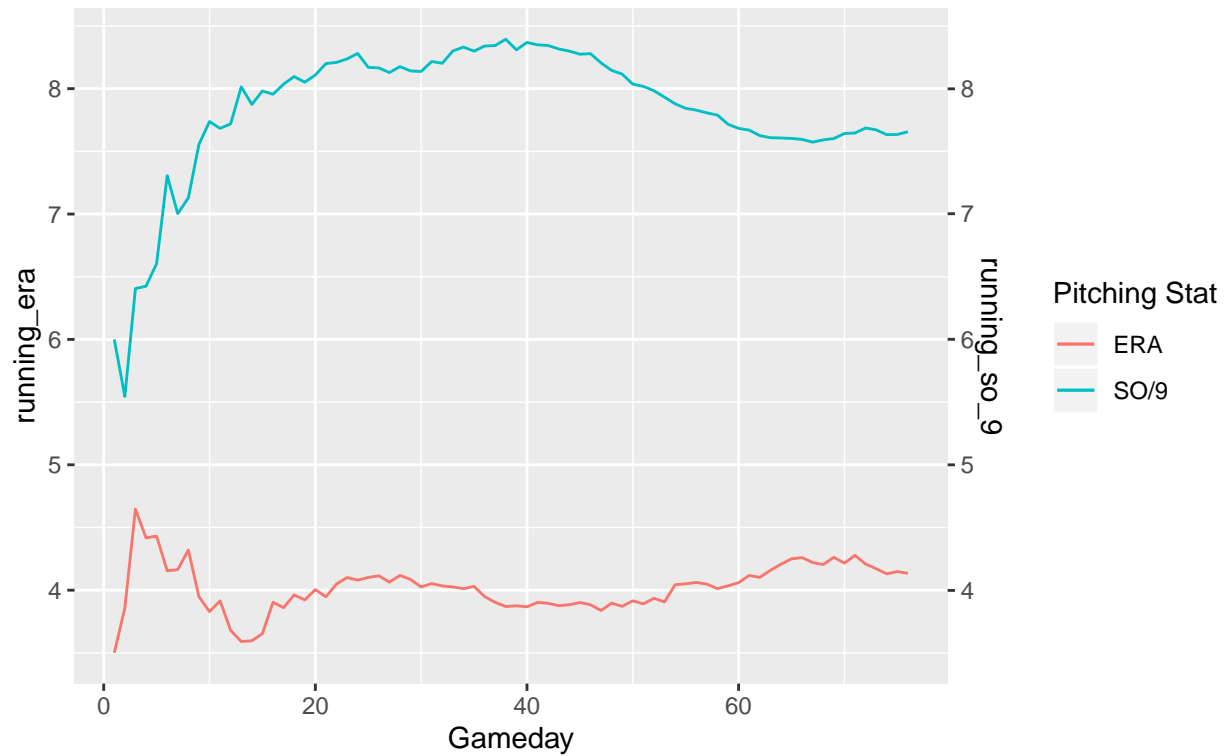

```r
write.csv(mario_data,'Mario_Baseball_Data_update.csv')
write.csv(leaguewide_data,'leaguewide_data.csv')
```